



# **Machine Learning, Parametrics, and Integrated Risk Analytics for Cost and Schedule**

**2021 Joint IT and Software Cost Forum**

**September 15, 2021**

GALORATH



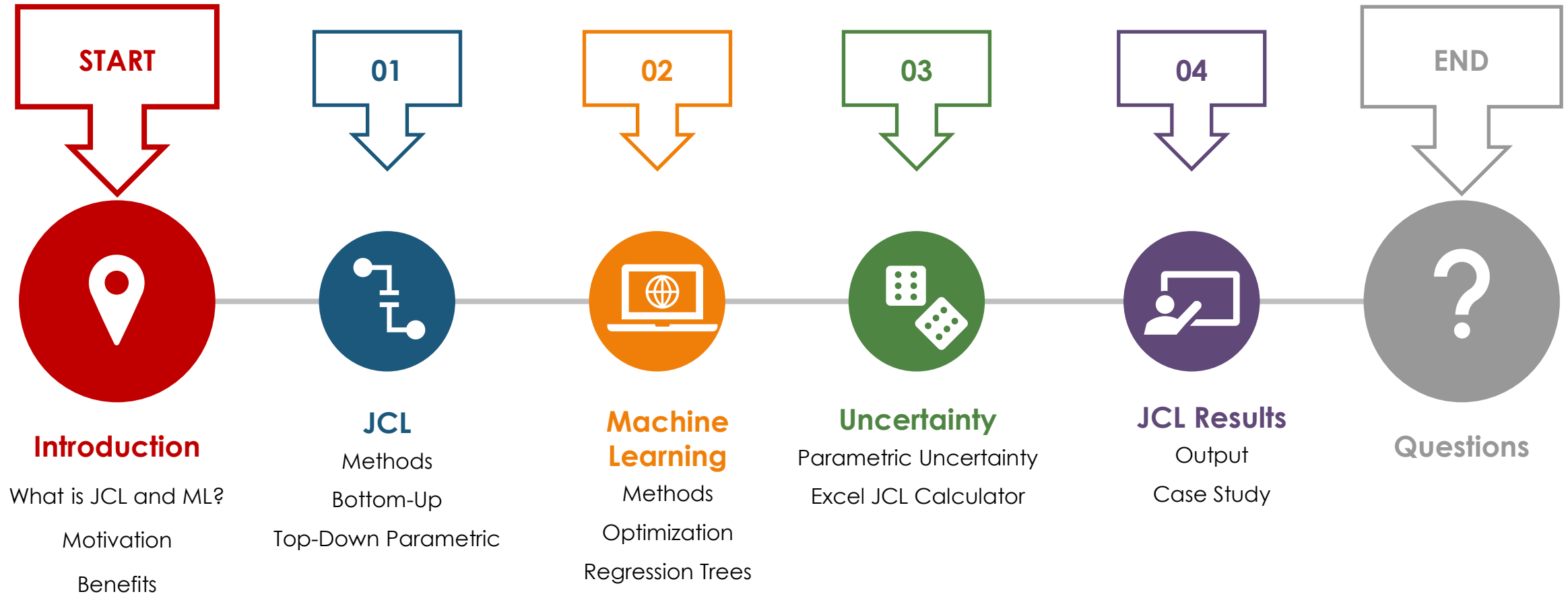
**We exist to empower informed decision making so that organizations can achieve their goals with greater confidence.**



G A L O R A T H



# Agenda





# Joint Confidence Level

## Holistic View of Uncertainty

An integrated uncertainty analysis of cost and schedule and a process combining a project's cost, schedule, and risk into a complete picture



## Joint Probability Approach

Represents a calculation combining the individual cost and schedule risk analyses into a joint probability distribution



## Integrated Cost and Schedule Risk

The goal is to identify the probability a given project or program cost will be equal to or less than the targeted cost and the schedule will be equal to or less than the targeted schedule



## More Robust

Because it is a more stringent requirement, the JCL is almost always higher than either the cost or schedule confidence



# Machine Learning



## Variety of Statistical Methods

Regression analysis is one tool of many in data science



## Traditional Methods

Includes regression analysis, clustering, dynamic programming, and numerical optimization



## Many Newer Methods

Decision trees, Deep Learning, Text Analytics, Reinforcement Learning



## Computationally Intense

Developments in machine learning take advantage of greater computing power

# Motivation

## COST AND SCHEDULE GROWTH

### A LEGACY OF DISASTER

	Olympics	Software/ IT	Dams	NASA/ DoD	Rail	Bridges/ Tunnels	Roads
Average Cost Growth	156%	43-56%	24-96%	52%	45%	34%	20%
Frequency of Occurrence	10/10	8/10	8/10	8/10	9/10	9/10	9/10
Frequency of Doubling	1 in 2	1 in 4	1 in 5	1 in 6	1 in 12	1 in 12	1 in 50
Average Schedule Delay	0%	63-84%	27-44%	27-52%	45%	23%	38%
Frequency of Schedule Delay	0/10	9/10	7/10	9/10	8/10	7/10	7/10

1

#### COMMON

Multiple Industries Experience Significant Cost and Schedule Growth – Has Been a Problem for a Long Time

2

#### FREQUENT

70-80% of Projects Experience Cost and Schedule Growth

3

#### HIGH

Cost: 50% or More on Average (Mean)  
Schedule: 30% or More on Average (Mean)

4

#### EXTREME (FOR COST)

Cost Growth in Excess of 100% Is a Common Occurrence in Most Projects (1 in 6)

# Track Record for Risk Analysis

## WORSE THAN RANDOM

Project	Cost Growth	Ratio of Actual Cost to 90% Confidence Level
1	0%	0.6
2	19%	1.1
3	31%	1.0
4	32%	1.1
5	greater than 45%	greater than 1.0
6	52%	1.5
7	84%	1.7
8	93%	1.6
9	121%	2.0
10	280%	2.2

*It's hard to improve if you don't know how well you have done in the past.*

1

### SCARCE

The results of risk analysis are rarely compared to the actual outcome – like a darts player that turns away from the board after throwing a dart

2

### WHAT LITTLE EXISTS IS NOT GOOD

The limited data available is mainly for cost  
The 90 percent confidence level means there is only a 10% probability that this level will be exceeded

3

### OPPOSITE OF EXPECTED

However, for the 10 risk analyses in the table, for only one was the actual cost less than the 90 percent confidence level

4

### EXTREMELY UNLIKELY

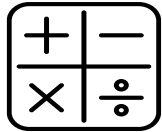
While a small data set, the odds of such an occurrence is extremely remote – 1 in 2.7 million  
You are more likely to be struck by lightning

**Poll Question: How many people/organizations track cost estimates all the way through to execution?**

# Motivation

## Collecting and Dealing with Data

Data are the foundation for sound estimates  
Bulk of time during analysis should be spent collecting, normalizing, and verifying data



### 1: NEED A LOT OF DATA

Collect as much data as possible

When limited data are available, consider the use of Bayesian methods to leverage all data available, including experience



### 2: CURSE OF DIMENSIONALITY

Amount of data needs grows exponentially in the number of variables

Often have more columns (variables) of data than rows (data points)



### 3: FEATURE EXTRACTION

Need to reduce the number of variables considered

Can leverage unsupervised learning techniques

# Benefits

## Joint Confidence Level

01

### Influences Decision Making

JCL provides a holistic view of the project in terms of possible outcomes given a program's level of risk and uncertainty

02

### Illuminates Correlation

Provides insight into correlation between cost and schedule

03

### Creates Project Plans

By providing an understanding of the relationship between cost and schedule, JCL helps to create and manage credible project plans

## Machine Learning

04

### Insightful

Unsupervised learning can provide more insight into your data, e.g., Clustering, K-Nearest Neighbors

05

### Predictive Accuracy

Techniques suited for categorical data, such as regression trees; ensembles; separates signal from noise; prevents overfitting

06

### Small Data Sets

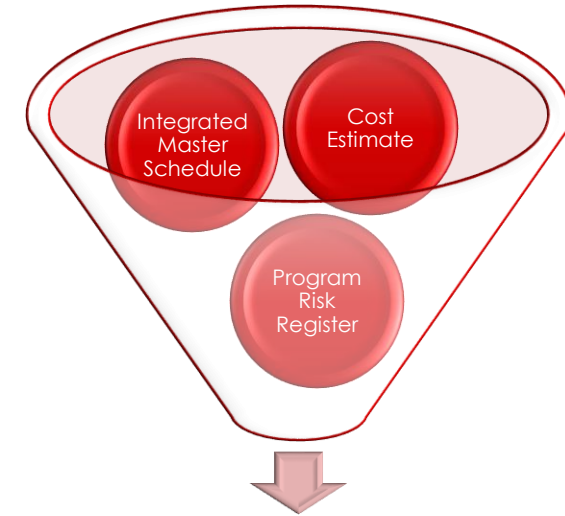
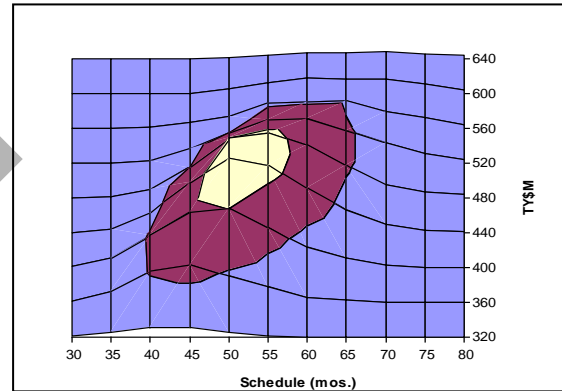
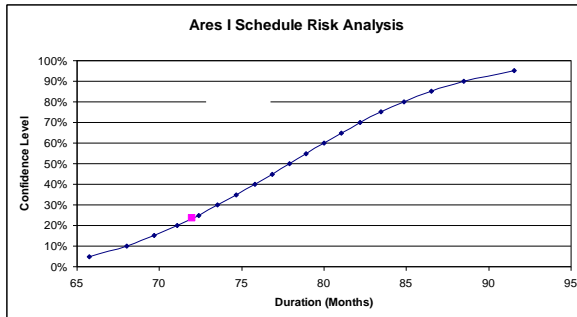
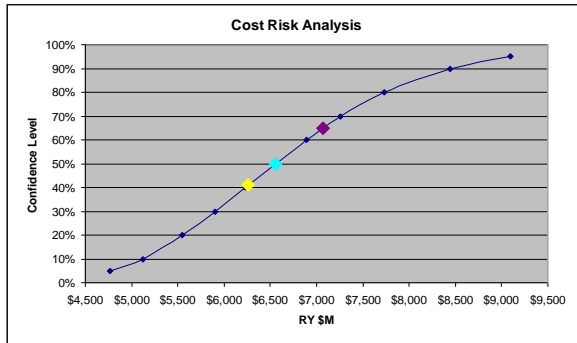
Use imputation to fill in holes in data; dimensionality reduction to deal with problem of more variables than data points; Bayesian methods to leverage other data and experience



# JCL Methods

# JCL Methods

## Top-Down vs Bottom-Up Comparison



**Integrated Risk Assessment**

### Top-Down

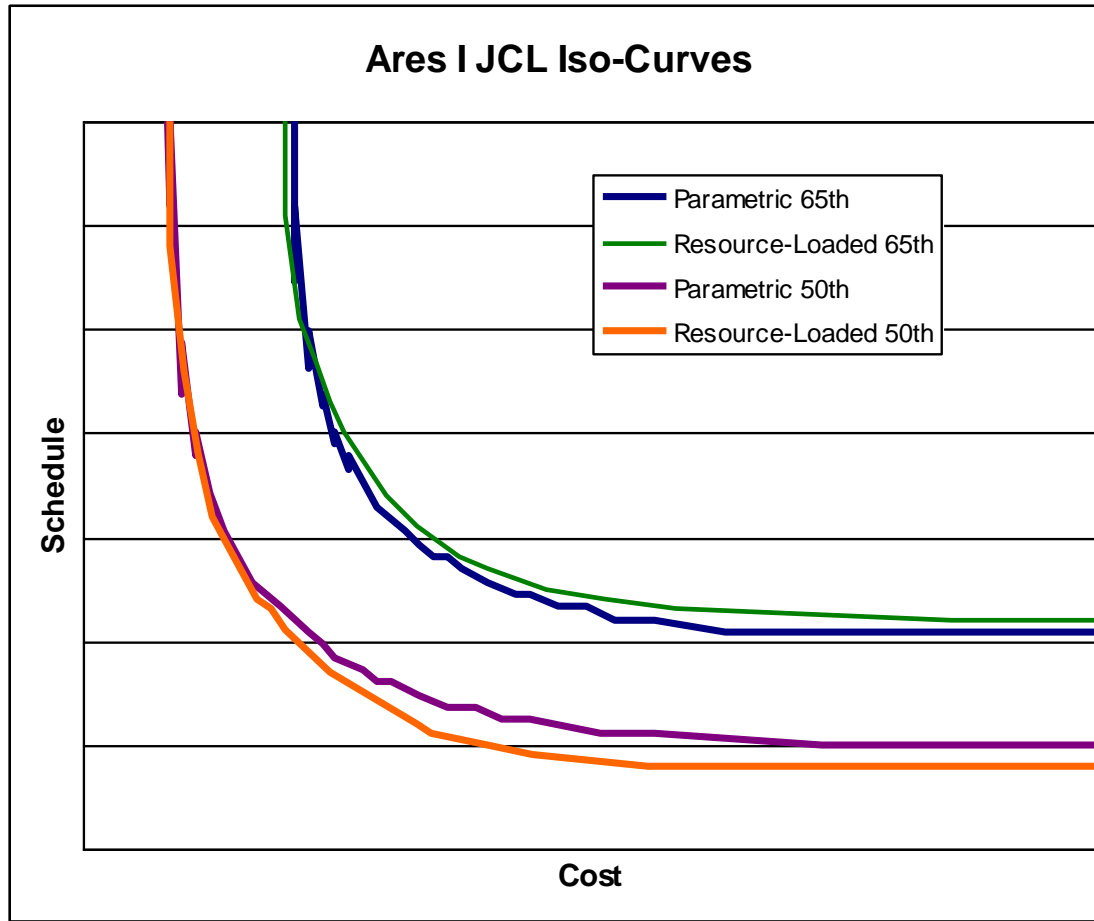
Uses parametric models to estimate cost and schedule uncertainty, then model the JCL as a joint probability distribution with assigned correlation. The successful NASA JCL was parametric; results can produce similar results

### Bottom-Up

Resource-loaded schedules are more time-intensive. A risk management system with a discrete and comprehensive risk list that captures cost and schedule risks is integrated into resource loaded schedule. (can be modeled in ACEIT software suite using JACS and MS Project)

# JCL Methods, Cont.

Top-down vs Bottom-Up Comparison



➤ **Accounting for Risk**

Tends to underestimate risk, easy to leave things out, plus it ignores unknown-unknowns, which are largely covered in the historical parametric data

➤ **Similar Results**

Results can produce similar results, as with the example on the left for the now-cancelled Ares I launch vehicle project

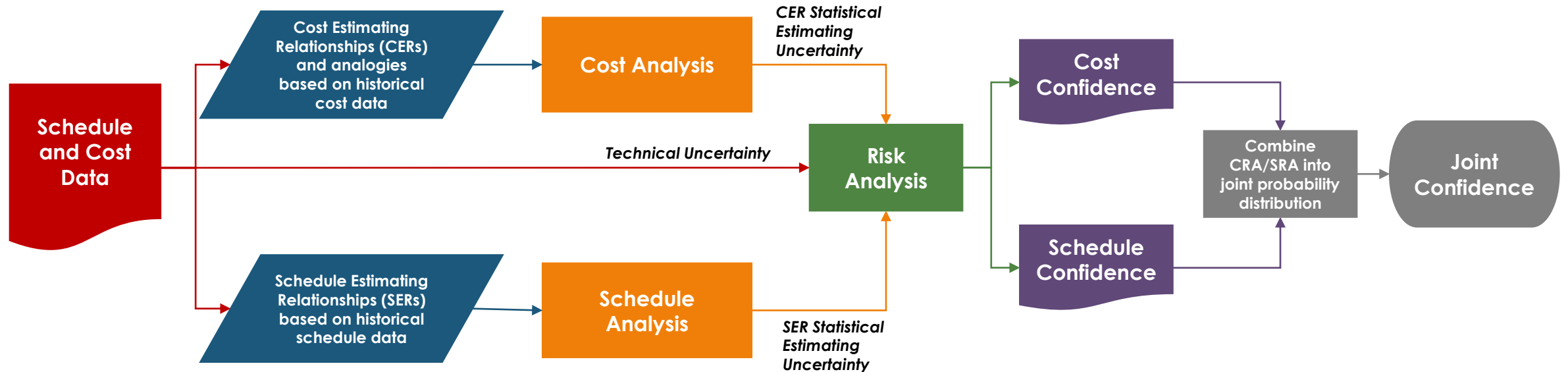
➤ **Agencies Using JCL**

NASA implements an agency JCL policy; the successful NASA JCL was parametric



# Top-Down Parametric Method

JCL Process



Step 1

Step 2

Step 3

Step 4

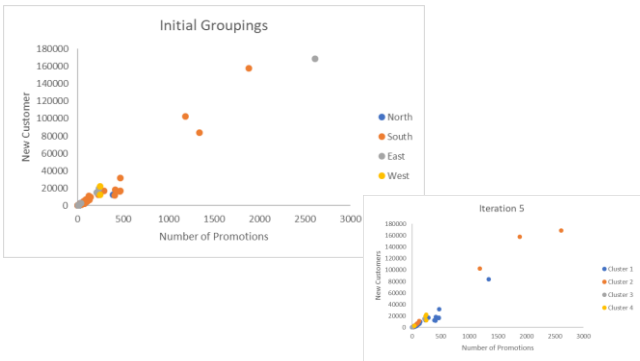
Step 5

Step 6

# Machine Learning

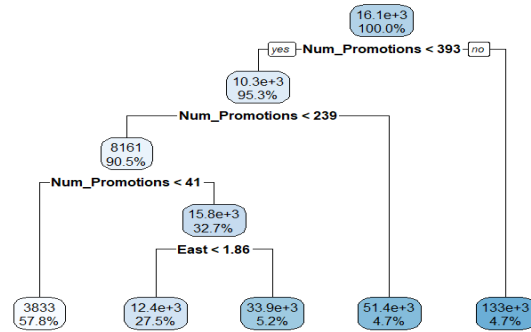
# Machine Learning Examples

Machine Learning techniques can be used to develop CERs and SERs to apply the top-down JCL parametric method



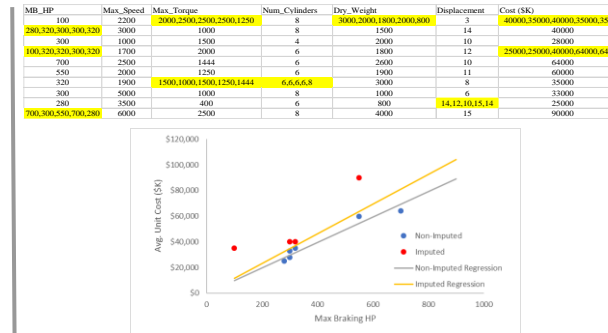
# Clustering

The practice of separating objects into similar groups and teases out relationships not easily discovered at first glance. Clustering begins with the random selection of centroids in the data then, iterative calculations are performed to optimize the positions of the centroids.



# Regression Trees

Can be used in preliminary data exploration to understand the most significant variables within a dataset. Since this method is non-parametric, it does not rely on data belonging to a particular type of distribution.



# Imputation

A powerful method useful for filling blanks when they are missing in a dataset. An analyst must understand the data intimately to know if a blank means that the factor is not applicable for that data point. Sometimes a blank does not reflect a nonresponse and should be observed “as is.”



# Natural Language Processing

Can be used to display the most frequently used words within the selected documents/websites. If interested in a specific topic, NLP can search for words related to this topic across mediums. A word cloud can be created which displays the most frequently used words within the select documents.



# Develop Cost Estimating Relationships (CERs) using Optimization

Optimization is a core component of Machine Learning

## What is Optimization?

Defined as a collection of mathematical principles and methods used for solving quantitative problems



## Basic Elements

Variables – free parameters the algorithm can tune

Constraints – boundaries within which the parameters must fall

Objective function – set of goals towards which the algorithm drives the solution

Almost all machine learning algorithms can be formulated as an optimization problem

## Goal

To minimize or maximize some function relative to some set, often representing a range of choices available in a certain situation. The function allows comparison to determine the “best” solution




## Application

Can be used to minimize or maximize a desired response related to endless problems within the Government realm or the private sector


# LINEST Function

## Optimal Solution

**Solver Parameters**


Set Objective:  

To: ☐ Max ☒ Min ☐ Value Of:

By Changing Variable Cells:  

Subject to the Constraints:

☐ Make Unconstrained Variables Non-Negative

Select a Solving Method:   Options

**Solving Method**

Select the GRG Nonlinear engine for Solver Problems that are smooth nonlinear. Select the LP Simplex engine for linear Solver Problems, and select the Evolutionary engine for Solver problems that are non-smooth.

Help
Solve
Close

Optimized  
coefficient  
values

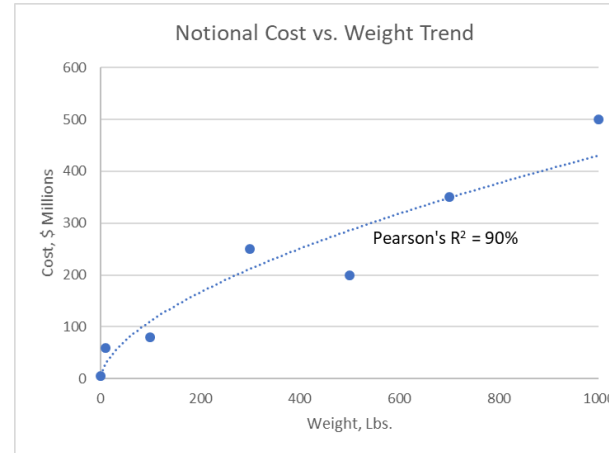
# The Trouble with Schedule Estimating Relationships (SERs)

SERs are generally more difficult to estimate using traditional parametric methods

1

## Regression Typically Works Well with Cost

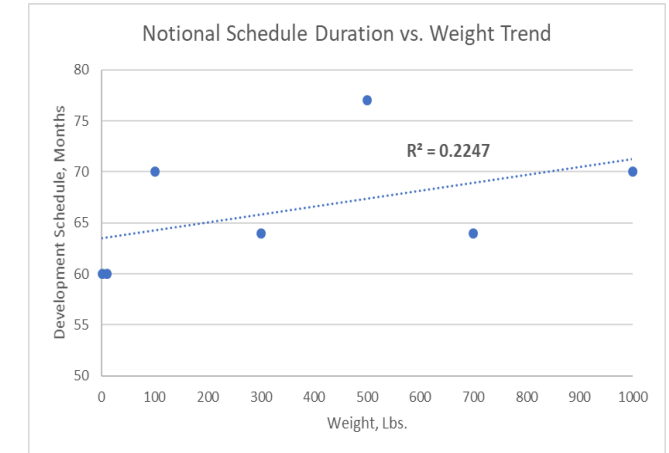
Wider spread in data points (cost and drivers) lends itself to meaningful trendlines



2

## Regression Often Does Not Work Well with Schedule

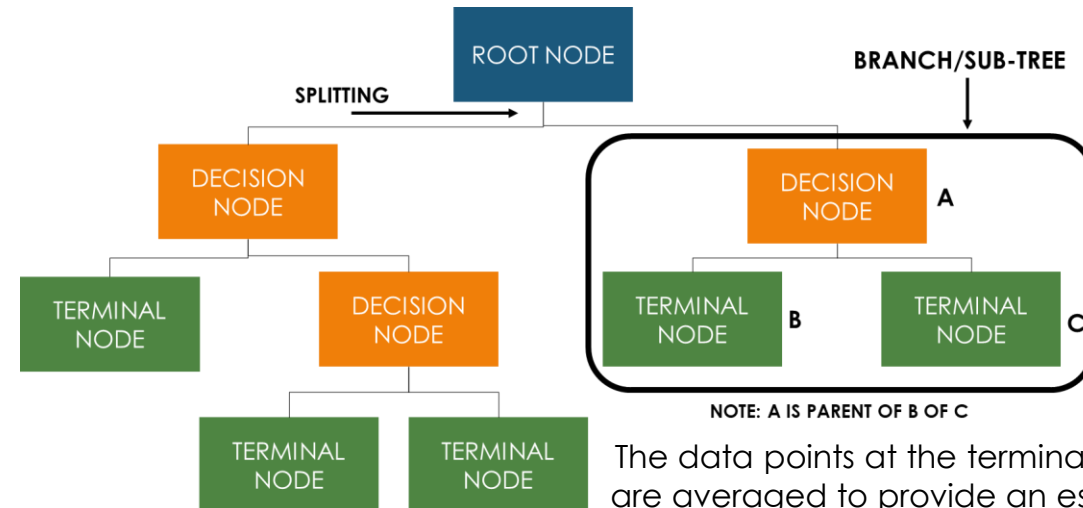
Schedule does not vary as much in terms of magnitude as cost – makes regressions less significant



3

## Machine Learning to the Rescue!

Alternative techniques such as regression trees are ideal when there is a significant amount of categorical data



4

## Regression Trees Can Work Better for Schedule

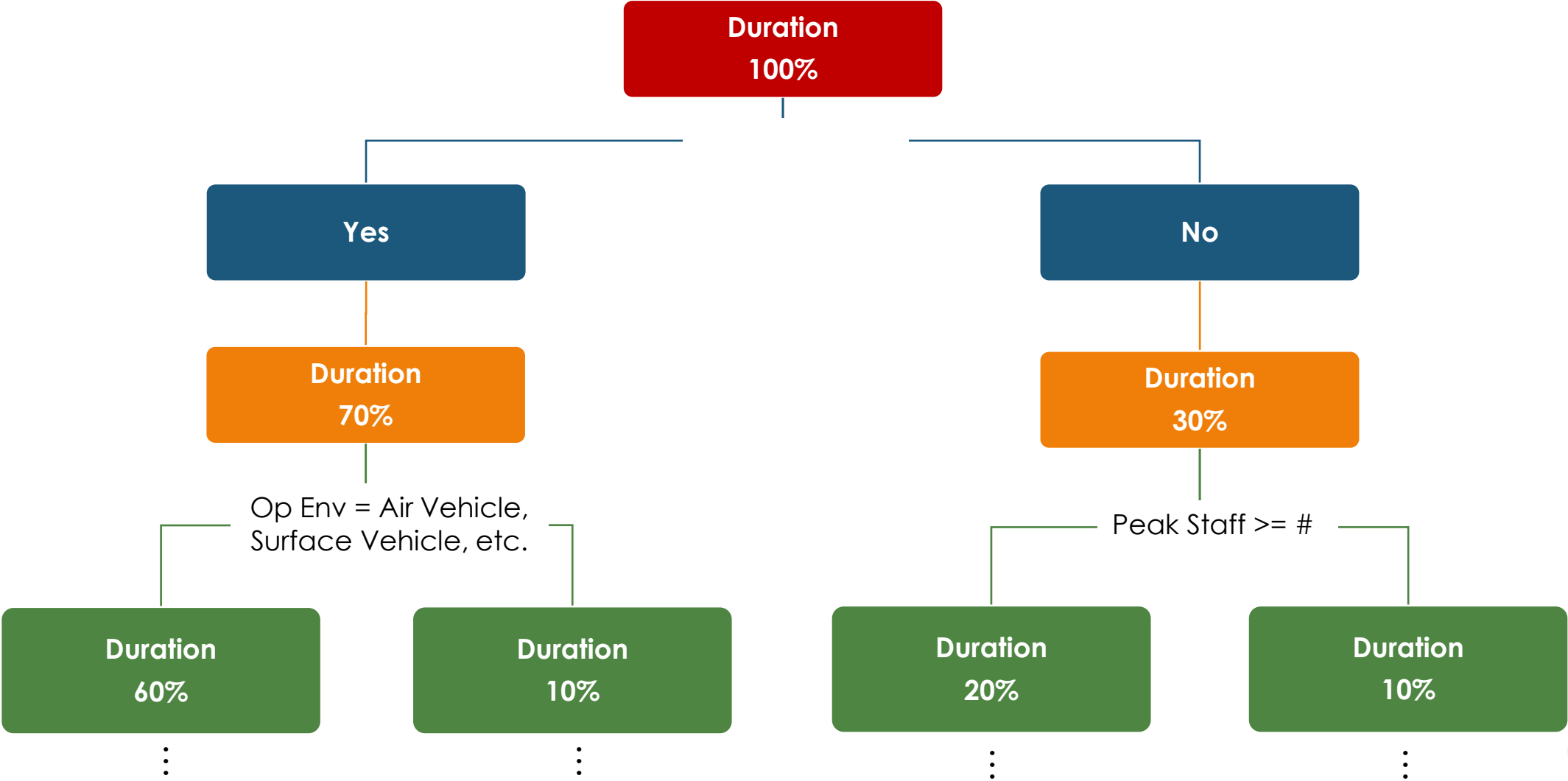
For software data use case, parametric schedule had no significant correlation with typical drivers; but with regression tree, Pearson's  $R^2$  was 50%

The data points at the terminal node are averaged to provide an estimate of the variables of interest based on the closest data points



# Schedule Estimating Relationship (SER) Regression Tree

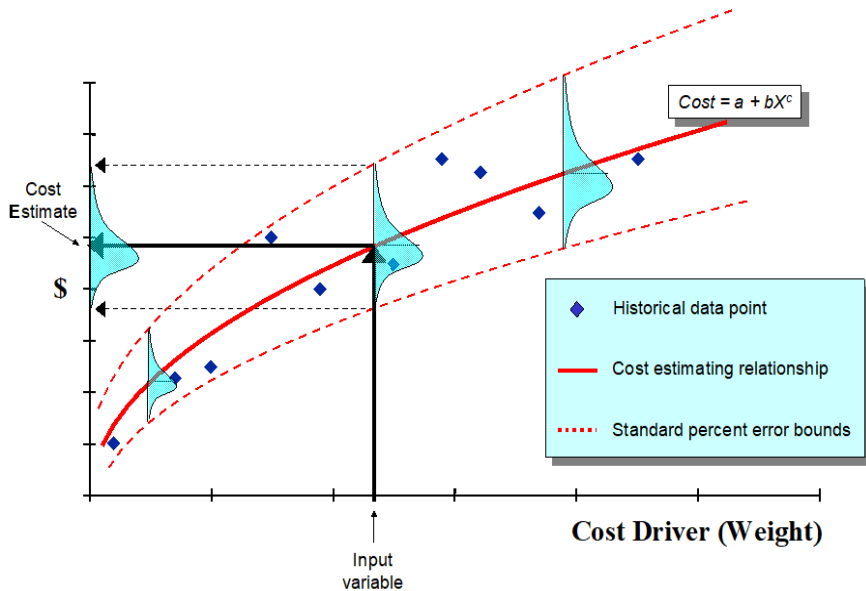
Software Program Example



# Uncertainty

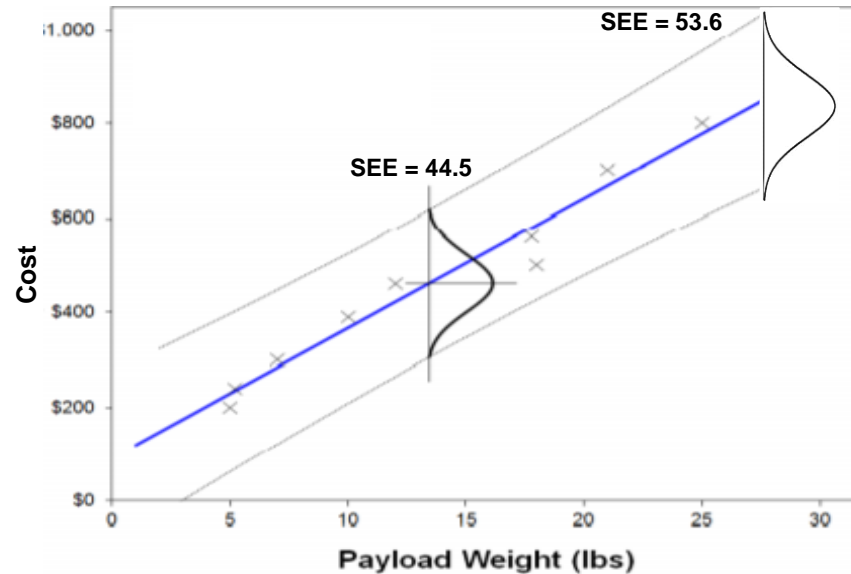
# CER and SER Estimating Uncertainty

Generate Cost and Schedule S-Curves by assigning uncertainty to the regression equations



## Parametric Uncertainty

Parametric CER/SER uncertainty represents uncertainty about the estimate's residual  $\varepsilon$ , (e.g.,  $Y = aX^b\varepsilon$ ). The further the input variable is from the center of mass data used to derive the CER/SER, the greater the uncertainty of the CER/SER



## CER/SER Uncertainty

The Standard Error of the Estimate (SEE) converts to a prediction interval to account for the distance of the estimate from the center of the CER/SER dataset. Uncertainty will increase (standard deviation gets larger) as the point estimate moves towards the data boundaries.

## Regression Equation

Regardless of the parametric method used, even if the independent variables are known precisely, the regression equation will return a result that is not certain

## Error Term

The error of the regression equation scales with the CER/SER result, making multiplicative error terms the preferred approach for modeling CER/SER uncertainty

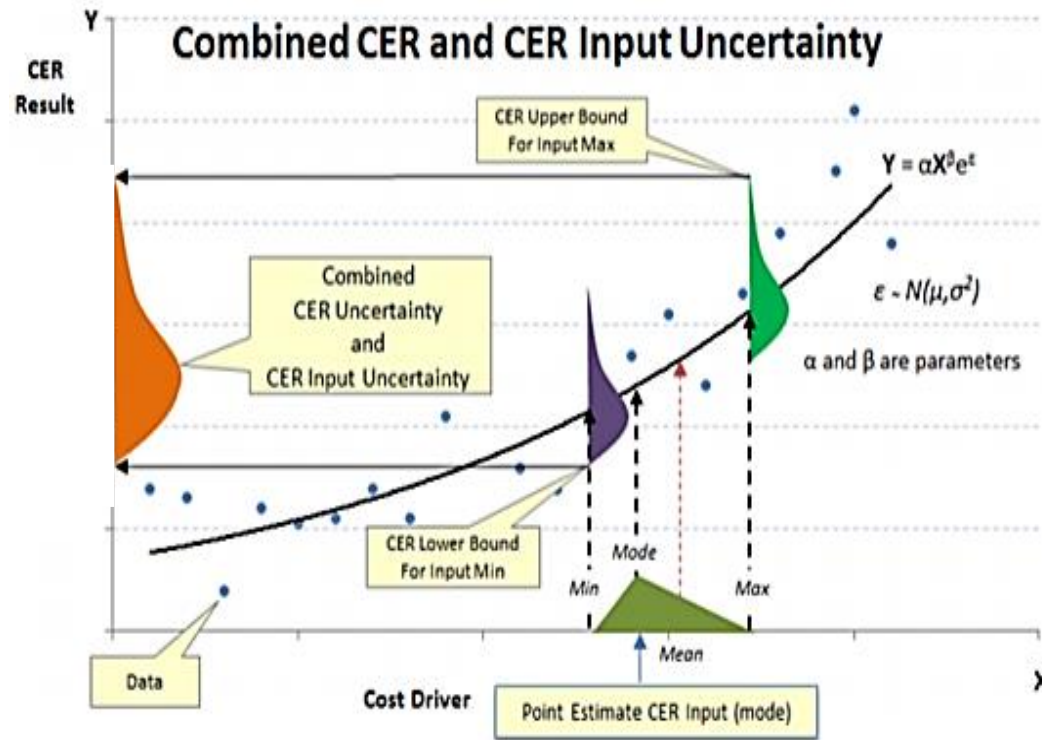
## Risk Parameters

The prediction interval or standard error provided by the regression analysis can be used to determine the CER/SER uncertainty bounds

# Combining CER

## Input and Estimating Uncertainty

Calculate uncertainty using propagation of errors



### 1 Input Uncertainty

Assume triangular distribution on input variables and run low (L), most likely (ML), and high (H) values through CER to obtain L, ML, H estimates. Calculate mean and standard deviation of triangular distribution.

$$\mu_x = \frac{L+ML+H}{3}$$

$$\sigma_x = \sqrt{\frac{L^2 + ML^2 + H^2 - L * ML - L * H - ML * H}{18}}$$

### 2 Estimating Uncertainty

Treated as the standard deviation of a lognormal distribution. Determine the mean and standard deviation in log space. Convert mean and standard deviation to unit space.

$$\mu_y = e^{\mu + \frac{1}{2}\sigma^2}$$

$$\sigma_y = \sqrt{(e^{\sigma^2} - 1)e^{2\mu + \sigma^2}}$$

### 3 Assumptions

Assume input and estimating uncertainty are independent; residuals are multiplicative

### 4 Propagation of Errors

Combine CER input (X) and estimating (Y) uncertainty by multiplying the means and standard deviations.

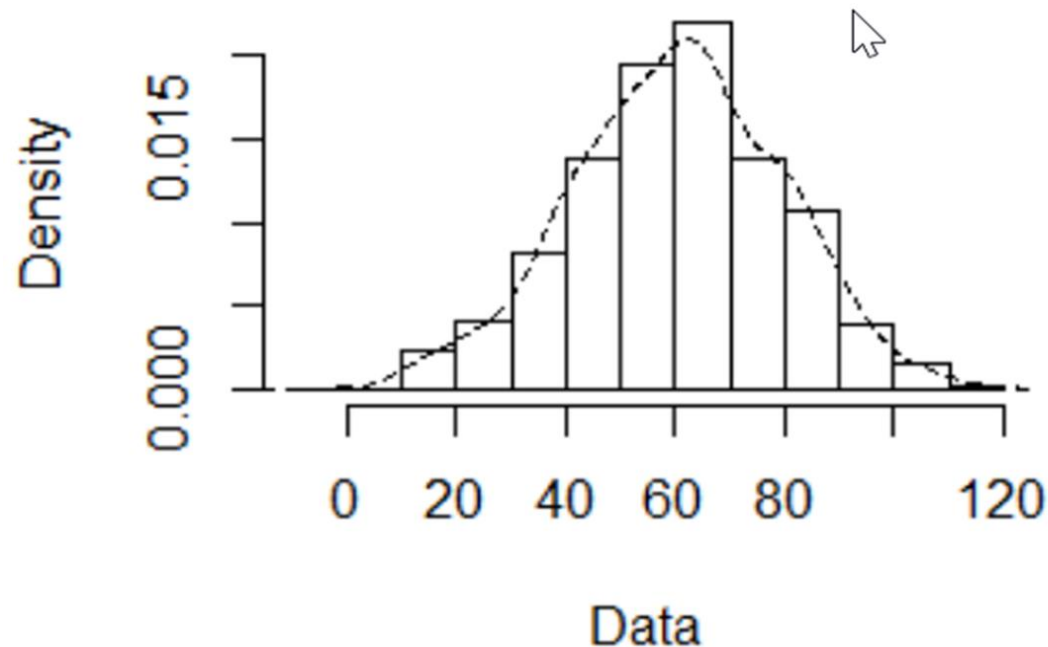
$$\mu(X * Y) = \mu_x * \mu_y$$

$$\sigma(X * Y) = \sqrt{\sigma_x^2 * \sigma_y^2 + \sigma_x^2 * \mu_y^2 + \sigma_y^2 * \mu_x^2}$$

# Regression Tree Uncertainty

Calculate total uncertainty using simulation

Simulation output fits a Gaussian/normal distribution based on sample mean and standard deviation of the simulation results



## 1 Input Uncertainty

Assume triangular distributions on SER inputs

## 2 Estimating Uncertainty

Regression tree residuals fit a Gaussian/normal distribution, determined mean and variance

## 3 Assumptions

Assume input and estimating uncertainty are independent; errors are additive

## 4 Simulation

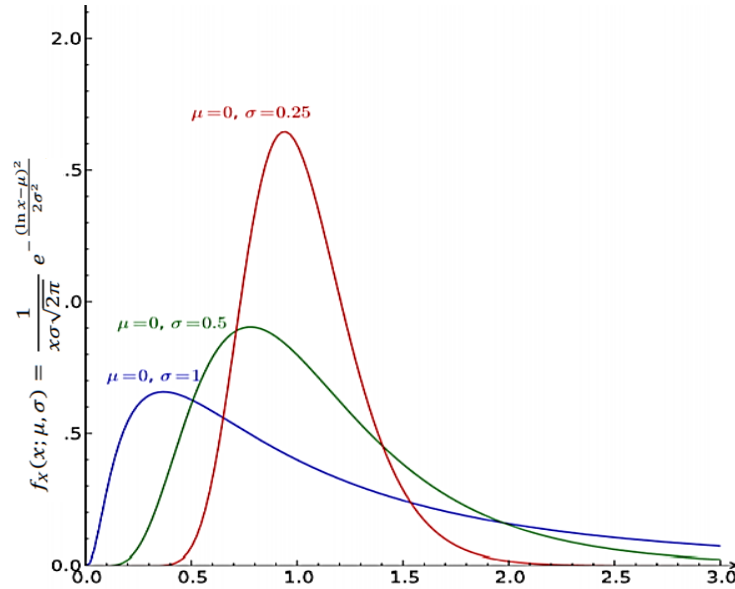
Conducted 1,000 trial simulation; on each draw:

1. Simulate the inputs, run the SER
2. Simulate a residual
3. Add the results of 1. and 2.



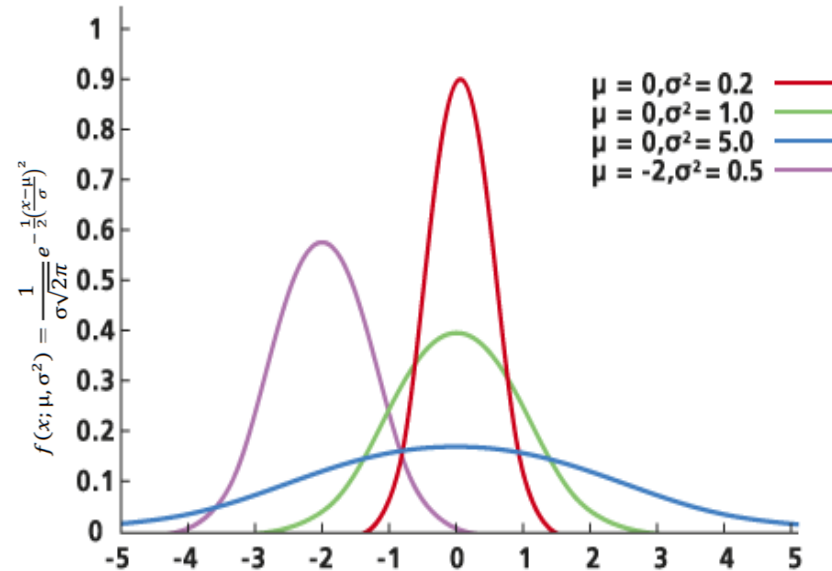
# Top-Down Parametric

After CER/SER is developed, conduct cost and schedule risk analyses



## Lognormal Distribution

Lower bound never less than zero and an upper bound of infinity. Probability is skewed right providing at least some probability of a large cost or schedule overrun.



## Normal Distribution

Unbounded in either direction with an equal probability of the low/high. Probability is not skewed where cost and schedule will more likely fall within the mean and extreme low and high values are less likely.

## Point Estimate

Determine the point estimate values for the project cost and project schedule

## Probability Distributions

To calculate a joint confidence, assume lognormal or normal risk distributions on cost and schedule using the mean and standard deviation as the parameters derived from the cost and schedule analyses

## Correlation

Assume linear correlation between cost and schedule based on historical data (e.g., Correlation value of 0.6 or 0.7)

# MS Excel Joint Confidence Level (JCL) Calculator

## Top-Down Parametric Method

Input mean values from cost and schedule risk analyses

Select probability distribution for cost and schedule

SECURITY WARNING

Macros have been disabled.

Enable Content

Select Enable Macros

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1		Mean	Sigma	Distribution	p	q	1st	99th								
2	Cost	530	159	Lognormal	6.2	0.3	256.4	739.5197								
3	Sched	45	6.75	Normal	3.8	0.1492	29.3	60.7								
4																
5	Correlation	0.6														
6	rho(1,2)	0.608411														
7																
11																
12																
13																
14																

Input a correlation value between cost and schedule

Input project budget and schedule values

Project Budget:

Project Schedule:

Your cost confidence is:

Your schedule confidence is:

Your project has joint cost and schedule confidence equal to:

\$600

40

71.5%

22.9%

21.7%

Calc. Joint Conf.

<---linear correlation between cost/schedule; **adjust** this value to attain rho correlation value you want for cost/schedule

<---rho is calculated value used for bivariate lognormal, and is the correlation value used in calculations

<---independent, not taking schedule into account

<---independent, not taking cost into account

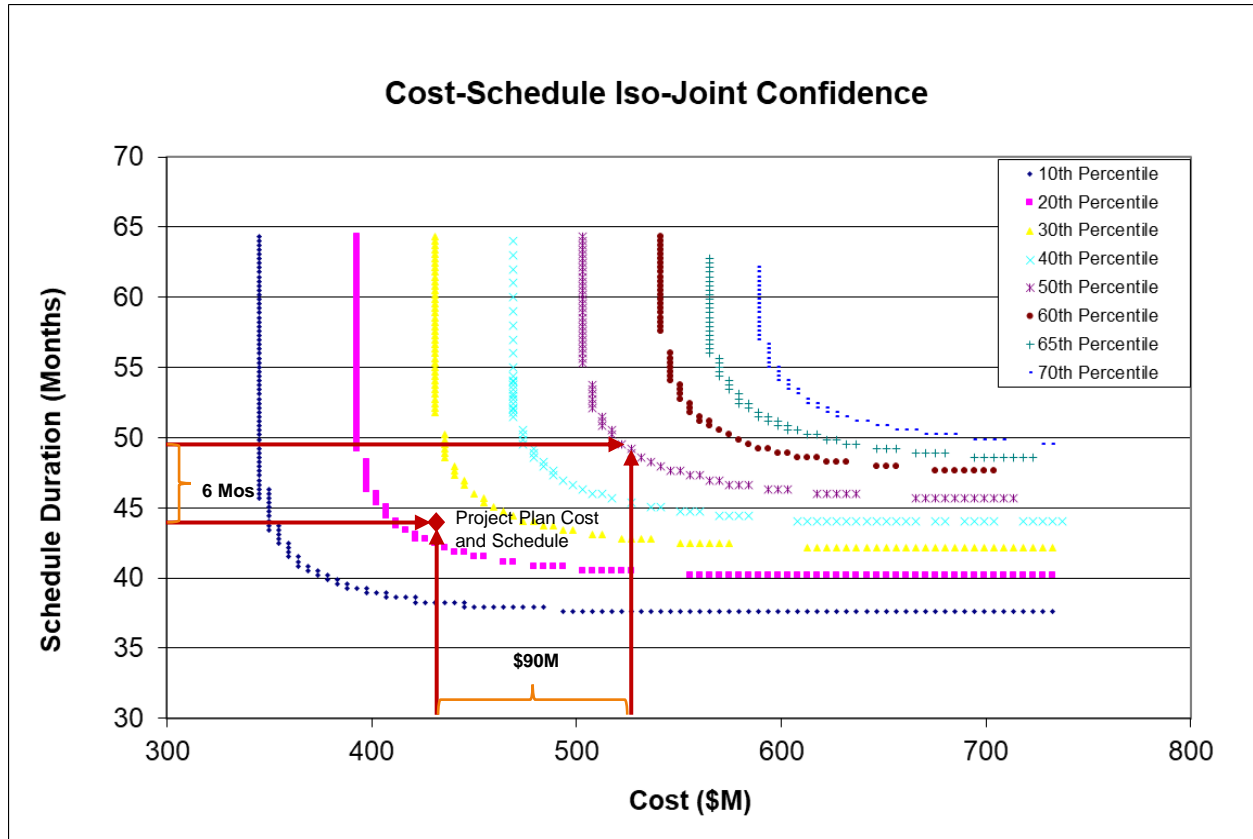
<---probability of achieving both project budget and project schedule



# JCL Results

# Top-Down JCL Result

JCL is calculated parametrically by combining the CRA and SRA into a joint probability distribution



Interpretation of JCL Result:

Current project plan is at 24% JCL (\$435M budgeted with a schedule of 43 months). **To achieve a 50% JCL, an additional \$90M in funding and 6 months of additional schedule is needed.**



## Project Plan JCL

Determine the JCL based on the current project plan to include the budget amount and schedule (months)



## Agency JCL Goals

Determine the agency's JCL requirement to establish a program budget to achieve the lifecycle cost and schedule



## Reserves

Depending on the relative importance of schedule vs. cost, determine the amount of cost reserves and additional schedule duration needed to achieve the agency's JCL goal

# Example Top-Down JCL Result

The JCL helps inform management of the likelihood of a project's programmatic success

Software Program Example

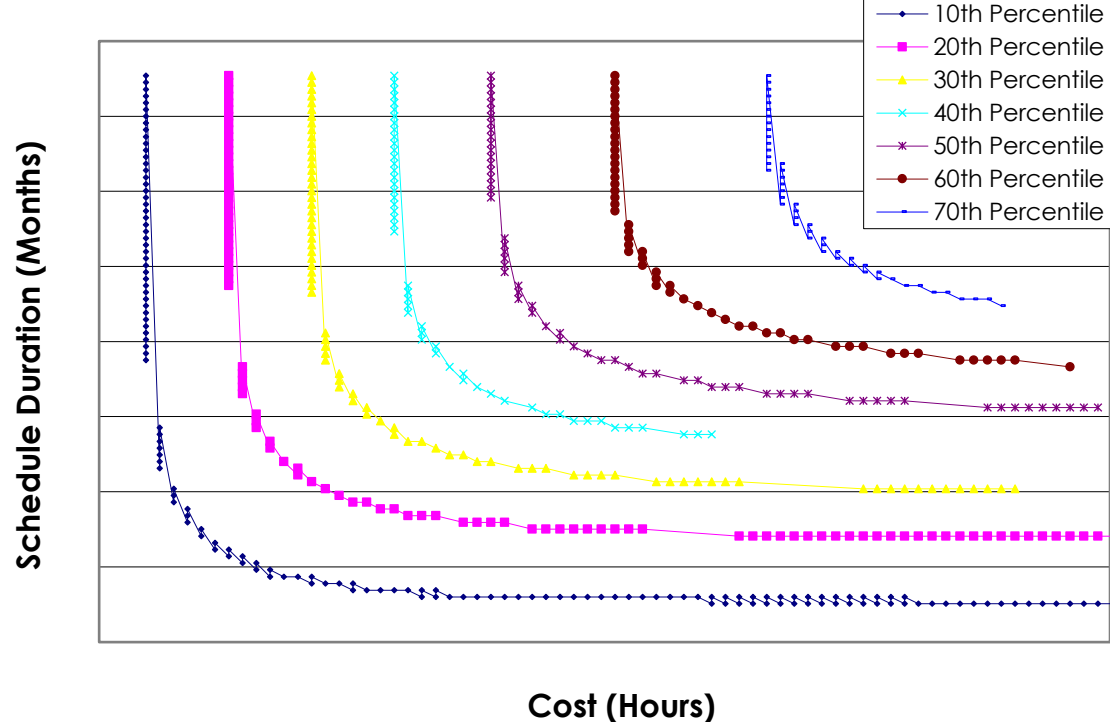
## Benefits of JCL to Management

➤ **Integrated Picture**  
JCL incorporates schedule, resources, and risk. Forces an integrated picture at the beginning and throughout the lifecycle

➤ **Forcing Function**  
NASA's JCL policy shows evidence that it has been an effective enforcer for better project management and executive decision making

➤ **Project Goals**  
JCL will help determine if results match project cost and schedule expectations. The agency's JCL objective will provide how much additional funding and schedule is needed to achieve the desired JCL

Software Program Example  
Cost-Schedule Iso-Joint Confidence

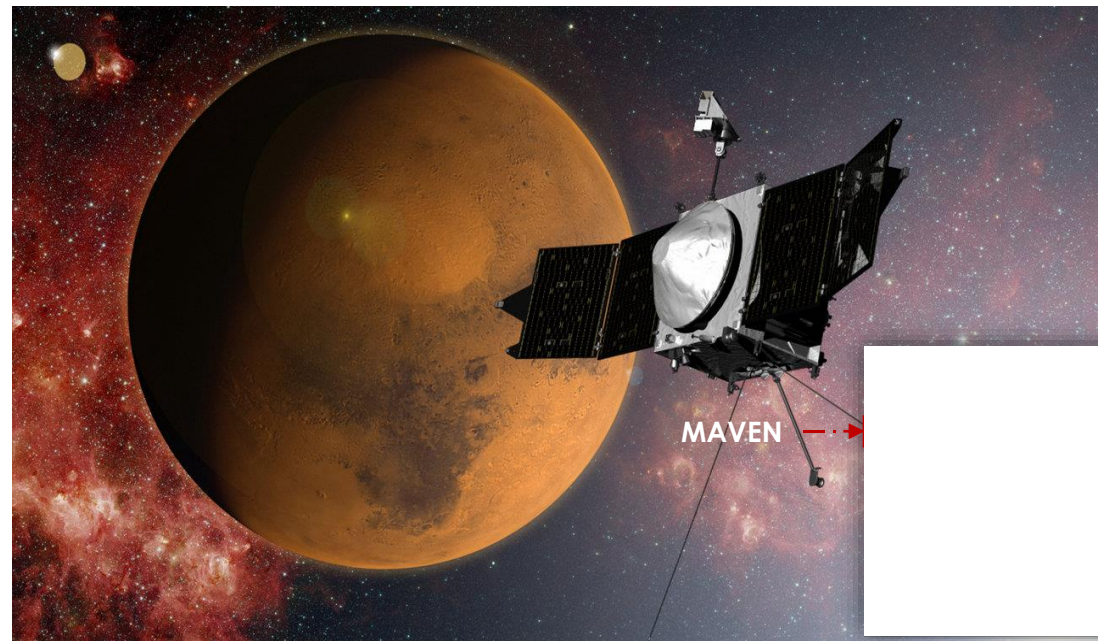
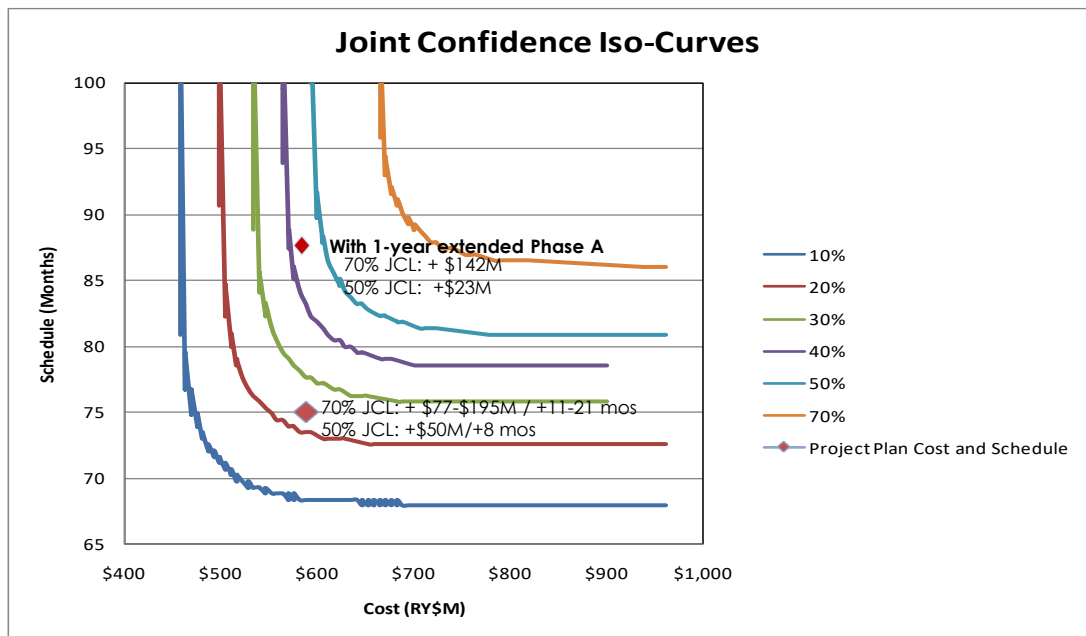




# Case Study: NASA MAVEN Spacecraft Program

## Success Story

Top-Down Parametric JCL Method



### JCL Estimate

In 2009, the MAVEN spacecraft program used the top-down parametric method to estimate the JCL. With the project plan cost and schedule, the JCL was estimated at 23% and if a year was added to the development schedule, the JCL was estimated at 44%.

### Program Actuals

In 2013, the actuals for cost and schedule for the MAVEN program came in at the 50% JCL estimated in 2009

# Presenters

Meet the Presenters



**Christian Smart**

**Chief Data Scientist**

Dr. Christian Smart is the Chief Data Scientist with Galorath. He is author of the forthcoming book *Solving for Project Risk Management: Understanding the Critical Role of Uncertainty in Project Management*. Dr. Smart is the VP for Professional Development with ICEAA. He regularly presents at conferences and has won several best paper awards. Dr. Smart received an Exceptional Public Service Medal from NASA in 2010 and has a PhD in Applied Mathematics.



**Sara Jardine**

**Senior Cost Analyst**

Sara Jardine is an experienced Operations Research Analyst who has worked directly for a broad variety of government agencies, including the Army, Navy, Veterans Affairs, and OUSD AT&L. She is skilled in Cost Management, Project Management, Requirements Analysis, Cost Analysis, Contract Management, and Budget Management. She has an MS in Project Management from The George Washington University and a BS in Mathematics from the University of Michigan.



**Kimberly Royce**

**Senior Data Scientist**

Kimberly Royce (CCEA®) is a Senior Cost Analyst for Galorath Federal. Starting her career as a Mathematical Statistician for the US Census Bureau, Kimberly transitioned to a career in Cost Analysis over 12 years ago. She has supported several Department of Defense hardware and vehicle programs. Kimberly earned an MS in Applied Statistics from Rochester Institute of Technology and a dual BS in Mathematics/Statistics from the University of Georgia.



# Q & A

## THE FUTURE. DELIVERED.

Galorath provides solutions that help organizational leaders make complex business decisions with confidence. Our predictive analytics products and services give complete insight into the implications of significant technical or financial decisions, allowing organizations to execute a plan with assurance and reach their goals with absolute certainty.

Learn more or schedule a demo  
(310) 906-6320 • [sales@galorath.com](mailto:sales@galorath.com)



Sara Jardine  
[sjardine@galorath.com](mailto:sjardine@galorath.com)

Kimberly Roye  
[kroye@galorath.com](mailto:kroye@galorath.com)

Christian Smart, PhD, CCEA  
[csmart@galorath.com](mailto:csmart@galorath.com)

